Conducting the NLM/AHCPR Large Scale Vocabulary Test: A distributed Internet-based experiment

Alexa T. McCray, May L. Cheh, Anantha K. Bangalore, Keyvan Rafei,
Amir M. Razi, Guy Divita, P. Zoe Stavri
National Library of Medicine
Bethesda, Maryland

The Large Scale Vocabulary Test, sponsored by the National Library of Medicine (NLM) and the Agency for Health Care Policy and Research (AHCPR), was conducted to determine the extent to which a combination of existing health-related terminologies cover vocabulary needed in health care information systems. The test was conducted over the Internet using a sophisticated World Wide Web interface with over 60 participants and over 40,000 terms submitted. This paper discusses the issues encountered in the design and execution of the experiment, including the design of the interface and the issues of recruitment, training, and guidance of remote participants over the Internet. Test data are currently undergoing expert review. Upon completion of the expert review, the results of the test will be fully reported.

INTRODUCTION

The goal of the Large Scale Vocabulary Test (LSVT) is to determine the extent to which a combination of existing health-related terminologies will be sufficient for the controlled vocabulary that is needed in a range of health care information systems, including electronic patient records. Health data of various types are increasingly being created, stored, and managed in electronic form. There is growing agreement that computerized health systems, in order that their data be comparable and more easily accessible, will require some level of language control [1]. Work on the Unified Medical Language System (UMLS) project over the past decade has resulted in the identification, incorporation, and linking of some thirty existing vocabularies through the UMLS Metathesaurus. The terminologies have been developed for a variety of purposes and are generally maintained by professional or specialty organizations, hospital or health care organizations, universities, international collaborative centers, or government agencies. Each of the vocabularies included in the Metathesaurus represents an effort by its developers to codify the major concepts of interest in the domain, together with a commitment to keep the terminology current.

To test the hypothesis that existing controlled vocabularies cover a significant portion of the terminology that is needed in health care information systems, we designed an experiment that involved testing the local terminologies of the participants against the set of vocabularies currently represented in the UMLS together with three planned additions [2]. The additions included those portions of SNOMED (The Systematized Nomenclature of Human and Veterinary Medicine) International that were not yet in the 1996 Metathesaurus, the British Read Clinical Classification, and the Regenstrief Institute's Logical Observations Identifiers, Names, and Codes.

To test the hypothesis that a large scale distributed experiment could be conducted successfully over the Internet, we designed and implemented a World Wide Web application for capturing submitted data, searching the Metathesaurus and its planned additions, browsing the UMLS information, and capturing users' decisions about the information returned by the system. The system was designed as a front end to the UMLS Knowledge Source Server [3]. Participation in the test was open to anyone who had a local terminology that was used to accomplish some task, for example, coding information in a hospital information system, providing access to patient data in a patient record system, or extracting and summarizing data for epidemiologic purposes. Further, it was assumed that participants had some interest in mapping their local terminologies to the UMLS. Technical requirements for participation included access to a good Internet connection together with the most current version of Netscape. The test was conducted over a five month period, from late August 1996 to mid January 1997. Test data are currently under expert review and the results of the test will be reported when the review is completed. This paper discusses the design and implementation of the LSVT interface and the issues encountered in conducting the experiment.

The experiment was conducted entirely over the Internet and was distributed over time and space. In

1091—8280/97/$5.00 © 1997 AMIA, Inc.

this respect it is an example of distributed research as it was envisioned by the founder of the World Wide Web [4]. Various groups have begun experimenting with using Web technology to conduct distributed research. Miller et al. [5] discuss the design of tools for carrying out collaborative research at two institutions. Shortliffe et al. [6] describe a research project involving six institutions. Their goal, in which they have had some initial success, is to share software, including data and knowledge interchange tools, as well as to conduct online distributed vocabulary building. The World Wide Web technology, with its standard protocols, its consistent interface, and, most importantly, its cross-platform client design, holds great promise for distributed research of all kinds. The Large Scale Vocabulary Test is an early example of the use of this technology to conduct a highly distributed experiment with participation by a large number of individuals over an extended period of time.

METHODS

Design and Development of the LSVT Interface

There were several considerations underlying the design of the vocabulary test system. While we were interested in having participants submit their local terminologies to the system, we were not interested in incorporating those terminologies directly in the UMLS. In other words, the goal was not vocabulary *building,* but, rather, it was to collect data to assess the extent to which health care terminology is already covered in existing vocabularies. This distinguishes our work from that of a number of other groups whose stated goal it is to build and even maintain a shared vocabulary system, e.g., [7,8]. A second design consideration was closely tied to the hypothesis that *multiple* existing vocabularies cover a significant portion of the needed terminology. Rather than searching different terminologies with different browsers and search engines and then comparing the relative merits of each of the terminologies, as was the case, for example, in the experiment described in [9], we designed a single interface to our UMLS Knowledge Source Server [3]. The Knowledge Source Server is based on concept, rather than term, matching and, therefore, searches through all constituent UMLS terminologies simultaneously. Third, because there is potentially quite extensive information in the UMLS about any given concept, we needed to consider the optimum type and amount of information to return when a user submitted a local term. Enough information needed to be displayed to allow the user to make a decision about the nature and correctness of the match made by the system. The tension between providing sufficient information for completing the task, on the one hand, and the desire to minimize the cognitive load on the user, on the other, involved extensive iteration in the implementation of the user interface. (See [10] for a good discussion of the need for incremental, iterative design in the development of interactive systems, and [11] for the results of a controlled experiment demonstrating the importance of paying attention to psychological factors in user interface design.)

The application we designed and implemented runs on a Web server with Netscape clients. It was a requirement of the test that all the participants use the latest version of Netscape (Netscape 3.0). This was necessary because the interface incorporated features such as frames, tables and javascript which were not available with earlier versions of Netscape or supported by other browsers at the time the test was conducted. The Web server communicates with the UMLS Knowledge Source (KS) Server through three back-end systems, one for the Metathesaurus data, one for the planned additions data, and the third for the approximate matching routines. The system modules are distributed over a Sparcserver 1000 and three Sparc Ultra 1 workstations. Figure 1 illustrates the system design.

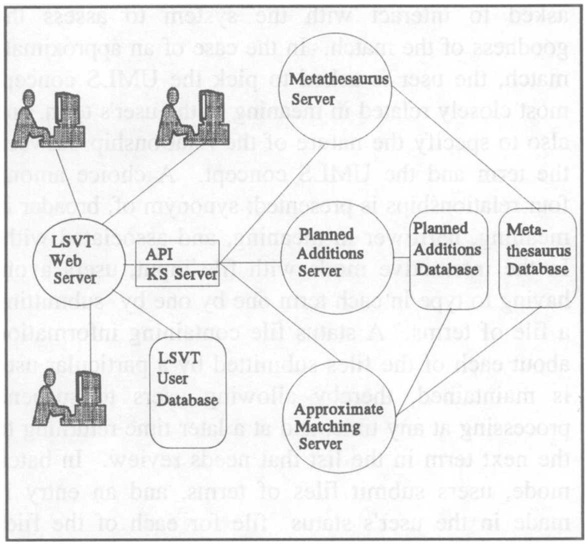

Figure 1: LSVT system architecture.

The Web client sends a request to the server; the request in this case is a term or a file of terms. The Web server, on receiving the request from the client, invokes a CGI program which opens a socket connection to the KS server. The term or terms are looked up first in the Metathesaurus server and then in the planned additions server. The system first looks for an exact match of the user's term with a UMLS concept. Exact matches are more than just string matches, however. The system uses the lexical normalization routines described in [3], and it maps user's terms to UMLS concepts, which in many cases include a large number of synonyms. Thus, the user's term "backache", for example, is an exact match with the Metathesaurus concept "back pain". When an exact match is made, the results, including the concept name, definition, semantic types, synonyms and ancestors are sent back to the CGI program, which then closes the socket connection, formats the results in HTML and sends them back to the client. If an exact match is not found, then a new connection is opened to an approximate matching server. This server computes a lexical distance measure which returns ranked lists of the ten most closely related concepts (with the best match at the top) from the Metathesaurus and another ten from the planned additions. The approximate matching algorithm is described in [12].

Users are able to interact with the system in three different modes, interactive, interactive with file input, and batch. In the interactive mode the user's terms are processed in real-time. In this mode, the user types in a term, the system looks for a match, and the user is asked to interact with the system to assess the goodness of the match. In the case of an approximate match, the user is asked to pick the UMLS concept most closely related in meaning to the user's term, and also to specify the nature of the relationship between the term and the UMLS concept. A choice among four relationships is presented: synonym of, broader in meaning, narrower in meaning, and associated with. In the interactive mode with file input, users avoid having to type in each term one by one by submitting a file of terms. A status file containing information about each of the files submitted by a particular user is maintained, thereby allowing users to suspend processing at any time, and at a later time returning to the next term in the list that needs review. In batch mode, users submit files of terms, and an entry is made in the user's status file for each of the files submitted. A scheduler process assigns each input file, depending on the number of terms in the file, to one of four queues. The scheduler also spawns four independent processes on three different machines to service the queues. These processes run in parallel, helping to achieve load balancing in the system. After each file is processed, an e-mail message is automatically sent to the owner of the file, indicating that the file is ready for review. At this point the user interacts with the results in much the same way as in the interactive modes, but all the system processing has already been done. In all cases users were given the opportunity to add information, including additional synonyms, a definition, and any comments about their term or about the UMLS match, as well any comments about the interface.

Test Procedures

Successful recruitment of testers was crucial to this experiment, and the usefulness of the study depended on participation from persons with tasks from a wide range of clinical applications. Adequate coverage of different areas of medicine and sufficient numbers of testers and terms were needed. Any organization or person who was willing to adhere to the test procedures, who could perform the test in the designated time frame, who had a good Internet connection, and who had a real task for which controlled vocabulary was desired was accepted into the experiment. To announce the test and to recruit potential participants, we used the Internet, a published journal article [2], a panel session at a national meeting, and a government solicitation for quotes. We created a demo version of the test interface and made it available on NLM's Web site inviting potential participants to try the demo and to register for the experiment. Most participants sent e-mail to us expressing their interest and were subsequently registered and invited to complete the required pretest.

Detailed instructions for both the pretest and the official test were available on the first screen of the LSVT interface. The pretest involved submitting 25 terms to the system in the available user interaction modes, stepping through all of the screens and answering questions just as they would in the official test, submitting their work together with any comments or questions they had about the test, and then waiting for e-mail from us in which we would clear up any misunderstandings about the test procedures and would let them know that they were free to go forward with the official test. Data from the pretests will not be included in the tabulation of the final results.

Data collected during the experiment included information about the participant, information about the participant's vocabulary, and information about the decisions made concerning the matches between the participants' terms and the concepts found in the UMLS. During the test, participants completed a term profile for each set of submitted terms. Term profile information included a description of the data task, the general purpose of the task, the care setting or facility, the specific type of care or specialty, and the specific segment of the patient record to which the controlled vocabulary applies.

Expert Review Process

Since the primary purpose of the test is to measure the extent to which existing terminologies match the vocabularies that are used in real health care tasks and not to measure the effectiveness of the test interface, a review process has been undertaken which involves participation by five content experts who hold either an M.D., Ph.D. or R.N. degree. The reviewer's job is to complete the task of finding matches that the interface has missed and to correct gross errors such as relationships presented in reverse order, or tester decisions that are otherwise clearly incorrect. Reviewers may use their expert knowledge of the medical field, their expert knowledge of the UMLS vocabulary, the UMLS Knowledge Source Server, and any other reference materials.

To accomplish the review task, another Web-based tool was developed. This tool has made it possible to involve individuals in different geographic locations (both East and West Coast) in the review process. Unique identifiers have been assigned to each term record, and reviewer-specific work lists are distributed regularly. The review interface was designed to address a number of requirements, specifically the need for access to participant provided term information, a listing of concepts matched to the term via the approximate matching algorithm, and access to online UMLS search tools. As each reviewer accesses the interface, the records from the work list are sequentially presented for consideration. Upon completion of the review of each term, the reviewer's decisions and comments are submitted and logged. Regular communication among the reviewers has included orientation meetings at the NLM, teleconferences, and exchanges of e-mail among all reviewers.

RESULTS

Ninety-five persons registered to participate in the experiment. Sixty-three, or two thirds of these, completed the pretest and participated fully in the official test. 73% of the test participants hold either an MD, RN, DDS or Pharmacy degree. The remaining 27% are PhD medical informaticians, medical librarians or medical students. Participants included organizations or individuals who were already funded as part of existing contracts with the NLM or AHCPR; organizations who were awarded a contract from the competitive request for quote specifically issued for this project; representatives of several professional societies who received a nominal amount to contribute vocabularies from their specialties; and many volunteers. Volunteers were either interested in contributing to the successful outcome of the experiment or were interested in using the collected data for their own purposes, or both. We provided all testers access to their own completed records. Thus, in addition to contributing to our experiment, participants were able to establish links between their local systems and the UMLS. Participants can use these connections to link to potentially useful additional information in the UMLS, including synonyms, definitions, semantic types, and hierarchical contexts.

During the five months of the experiment more than 40,000 terms were submitted. The number of terms submitted per tester ranged from under one hundred to several thousand. Figure 2 shows the geographic distribution of the test participants.

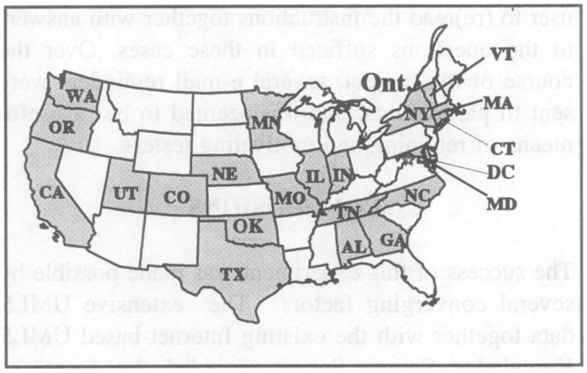

Figure 2: Test participants contributed terminology from 21 states, the District of Columbia, and Ontario, Canada.

Although the experiment proceeded smoothly for most participants, a few users did experience some technical difficulties. On a few occasions the NLM server was down for maintenance or hardware and software updates, including some bug fixes. A handful of users experienced down time because they had used special characters in the terms or filenames they submitted. We solved the problem of the special characters in the terms but asked users to avoid using certain characters in their filenames. Another problem involved a user typing ahead of the transmission during very slow network times. This was a rare occurrence but, nonetheless, confusing. A source of frustration to some participants who used PC's was that the LSVT interface did not fit comfortably on their screens, resulting in the need to do a fair amount of scrolling to read all of the information presented. This problem was easily remedied by changing the font size on the user's Netscape client; however a few users became quite frustrated before contacting us. Data from all but one user were successfully collected over the Internet using the LSVT interface. The one exception was the tester who was caught in the serious access problems faced by America Online users during December 1996 and January 1997. After repeated tries and mounting frustration levels, this tester was asked to submit her terms and decisions on paper. These were subsequently entered into the system by the project team.

Communication with participants included e-mail messages and telephone calls. Questions included clarifications for vocabulary content, comments concerning terms or the interface, and reporting of technical difficulties. Occasionally, the nature of the e-mail message indicated that the participant had not read the instructions. A brief e-mail reminding the user to (re)read the instructions together with answers to the questions sufficed in these cases. Over the course of the project, several e-mail reminders were sent to participants, and this seemed to be a useful means of retaining and motivating testers.

CONCLUSIONS

The success of this experiment was made possible by several converging factors. The extensive UMLS data together with the existing Internet-based UMLS Knowledge Source Server provided the necessary content and computational platform to design the LSVT application. The World Wide Web and the increasing capabilities of Netscape and related browsers allowed us to design a sophisticated tool that was successfully used over an extended period of time by individuals in many different locations. The interest in the community in matters related to controlled health care terminologies meant that we were able to attract many individuals to participate in the test and resulted in large numbers of terms being submitted. Following the completion of the expert review, the test data will be fully analyzed and reported. The results of the analysis should help evaluate the potential of exisiting health care vocabularies to meet the full range of clinical, public health, and research requirements.

REFERENCES

1. Board of Directors of the American Medical Association. Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. JAMIA, 1994, 1(1):1-17.
2. Humphreys BL, Hole WT, McCray AT, Fitzmaurice MJ. Planned NLM/AHCPR large-scale vocabulary test: Using UMLS technology to determine how well controlled vocabularies cover terminology needed for health care and public health. JAMIA, 1996, 3(4):281-287.
3. McCray AT, Razi AM, Bangalore AK, Browne AC, Stavri PZ. The UMLS Knowledge Source Server: A versatile Internet-based research tool. Proc Annu Symp Comput Appl Med Care, 1996, 164-168.
4. Berners-Lee T, Cailliau R, Luotonen A, Nielsen HF, Secret A. The World Wide Web. Communications of the ACM, 1994:37(8):76-82.
5. Miller PL, Nadkarni PM, Kidd KK, et al. Internet-based support for bioscience research: a collaborative genome center for human chromosome 12. JAMIA, 1995, 2(6):351-64.
6. Shortliffe EH, Barnett GO, Cimino JJ, Greenes RA, Huff SM, Patel VL. Collaborative medical informatics research using the Internet and the World Wide Web. Proc Annu Symp Comput Appl Med Care 1996, 125-129.
7. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a controlled medical vocabulary server: The VOSER project. Computers and Biomedical Research, 1994:27(6):472-507.
8. Gennari JH, Oliver DE, Pratt W, Rice J, Musen MA. A Web-based architecture for a medical vocabulary server. Proc Annu Symp Comput Appl Med Care 1995, 275-9.
9. Campbell JR and Payne TH. A comparison of four schemes for codification of problem lists. Proc Annu Symp Comput Appl Med Care, 1994, 201-5.
10. Rosson MB, Maass S, Kellogg WA. The designer as user: Building requirements for design tools from design practice. Communications of the ACM, 1988:31(11): 1288-1298.
11. Poon, AD, Fagan LM, Shortliffe EH. The PEN-Ivory Project: Exploring user-interface design for the selection of items from large controlled vocabularies of medicine. JAMIA, 1996:3(2): 168-183.
12. Aronson AR, Rindflesch TC, Browne AC. Exploiting a large thesaurus for information retrieval. Proceedings of RIAO, 1994, 197-216.